Int J Life Cycle Assess (2013) 18:185–192
DOI 10.1007/s11367-012-0454-3

Sampling error in US field crop unit process data for life cycle assessment

Joyce Smith Cooper · Ezra Kahn · Robert Ebel

Received: 8 September 2011 / Accepted: 6 June 2012 / Published online: 21 June 2012

Abstract
Purpose The research presented here was motivated by an interest in understanding the magnitude of sampling error in crop production unit process data developed for life cycle assessments (LCAs) of food, biofuel, and bioproduct production. More broadly, uncertainty data are placed within the context of conclusive interpretations of comparative bioproduct LCA results.
Methods Data from the US Department of Agriculture's Agricultural Resource Management Survey were parameterized for 466 crop–state–year combinations, using 146 variables representing the previous crop, tillage and seed operations, irrigation, and applications of synthetic fertilizer, lime, nitrogen inhibitor, organic fertilizer, and pesticides. Data are described by Student's *t* distributions representing sampling error through the relative standard error (RSE) and are organized by the magnitude of the RSE by data point. Also, instances in which the bounds of the 95 % confidence intervals are less than zero or exceed actual limits are identified.
Results and discussion Although the vast majority of the data have a RSE less than 100 %, values range from 0 to 1,600 %. The least precision was found in data collected between 2001 and 2002, in the production of corn and soybeans and in synthetic and pesticide applications and irrigation data. The highest precision was seen in the production of durum wheat, rice, oats, and peanuts and in data representing previous crops and till and seed technology use. Additionally, upwards of 20 % of the unit process, data had 95 % confidence intervals that are less than or exceed actual limits, such as an estimation of a negative area or a portion exceeding a total area, as a consequence of using a jackknife on subsets of data for which the weights are not calibrated explicitly and a low presence of certain practices.
Conclusions High RSE values arise from the RSE representing a biased distribution, a jackknife estimate being nearly zero, or error propagation using low-precision data. As error propagates to the final unit process data, care is required when interpreting an inventory, e.g., Monte Carlo simulation should only be sampled within the appropriate bounds. At high levels of sampling error such as those described here, comparisons of LCA bioproduct results must be made with caution and must be tested to ensure mean values are different to a desired level of significance.

Keywords Error · Inventory data · Life cycle assessment · Meta data · Parameterization · Uncertainty

Responsible editor: Andreas Ciroth

Electronic supplementary material The online version of this article (doi:10.1007/s11367-012-0454-3) contains supplementary material, which is available to authorized users.

J. S. Cooper (✉) · E. Kahn
Design for Environment Laboratory, University of Washington,
Box 352700, Seattle, WA 98195-2600, USA
e-mail: cooperjs@u.washington.edu

R. Ebel
Economic Research Service, US Department of Agriculture,
355 E Street SW,
Washington, DC 20024-3221, USA

1 Introduction

For life cycle assessment (LCA), ISO 14044 defines uncertainty analysis as a "systematic procedure to quantify the uncertainty introduced in the results of a life cycle inventory analysis due to the cumulative effects of model imprecision, input uncertainty and data variability" and notes that "either ranges or probability distributions are used to determine

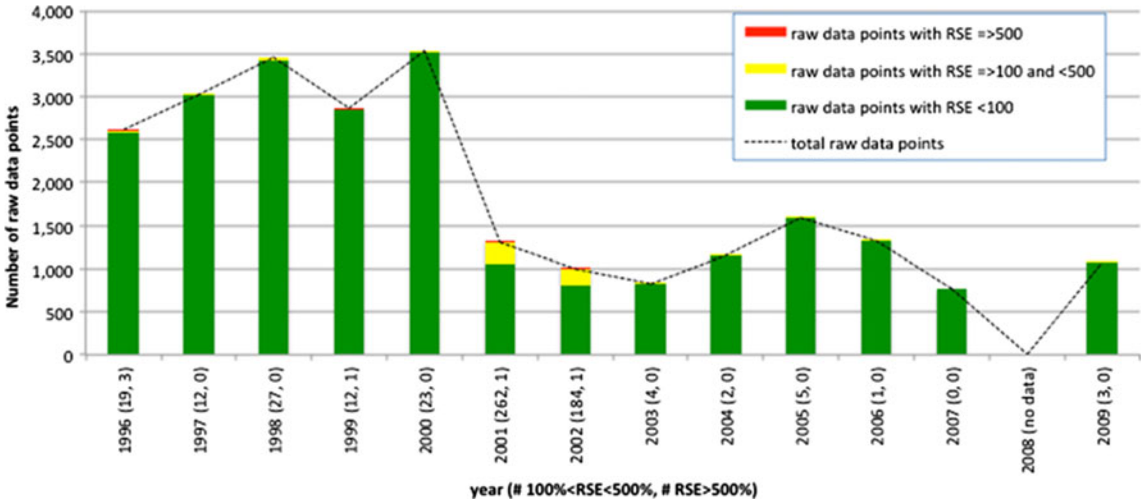

Fig. 1 ARMS raw data by year

uncertainty in the results." However, the vast majority of LCAs do not consider data variability, in part because of a lack of variability estimates, e.g., in LCA databases. One exception lies in data put forth by the ecoinvent Centre,[1] which uses qualitatively derived data quality scores to estimate the "additional" uncertainty resulting from lower data quality as the "square of the geometric standard deviation (95 % interval—SDg95)" (Weidema and Wesnæs 1996). However, Lloyd and Ries (2007) warn that unless distribution forms and parameters are defined for specific scores and parameter contributions, there is no basis for their accuracy. Noting that the ecoinvent Centre has commissioned an empirical study to validate and revise the basic uncertainty factors used in the estimation of the SDg95 (Weidema et al. 2011), here, we consider data variability outside of this "additional" uncertainty.

Consider for example sampling error, a measure of the inaccuracy caused by observing a sample instead of an entire population. In an LCA, data might be developed based on the operation of a single or multiple industrial sites sampled over some timeframe, or they might be estimated using a computational model that quantifies production as a function of a sample of feedstock compositions (e.g., the composition of crude oil or a bio-feedstock). Basic statistics provide methods for using such sample data to estimate probability distributions (functions that describe the probability that a random variable will take certain values, such as normal, Student's *t*, lognormal, Poisson, and Bernoulli distributions, etc.) for use in uncertainty analysis in an LCA. Further, the characteristics of the data and the sampling method dictate the appropriateness of distribution form; e.g., whereas a normal distribution might be used at large sample sizes, Student's *t* distribution can better represent a population based on smaller sample sizes by increasing the probabilities at the extremes of the distribution (i.e., the tails are larger than in a normal distribution).

As the use of LCA in the development of public policy and law (e.g., in the USA, the 2007 Energy Independence and Security Act) and in the comparison of products (e.g., in the development of Product Category Rules) is rising, it seems data uncertainty analysis based on well-developed statistical methods will be demanded from LCA practitioners. Questions that immediately arise relate to the magnitude of variability in the data being used in LCA, irrespective of the consideration of the "additional" data quality-based uncertainty. Specifically, is the variability of LCA data small or large as compared to mean exchange values, and can we conclusively interpret comparative LCA results?

Consider for example a comparison of the life cycles of a conventional fuel and a biofuel in which the conventional fuel has an estimated mean greenhouse gas emission of 47 g CO_2e/MJ and the biofuel of 38 g CO_2e/MJ. Without consideration of variability, the biofuel is found superior to the conventional fuel, offering a 20 % improvement. If the relative standard errors (the RSEs,[2] also called coefficients of variation) are, e.g., 5 and 10 % for the conventional fuel and biofuel, respectively, and in both cases, 30 random samples were taken from much larger populations that are assumed to be normal, at a significance level of 5 %, the means are found to be significantly different using a two-sample *t* test. In this case, drawing the conclusion that the biofuel is superior is valid. Alternatively, under the same sampling scheme and at the same significance level, if the

[1] Available at http://www.ecoinvent.ch/

[2] The RSE is the standard error (SE) of the mean divided by the mean and expressed as a percentage. Because the SE is the sample standard deviation divided by the square root of the sample size, the RSE is intended to represent the difference between the estimate and the true value with respect to the magnitude of the mean.

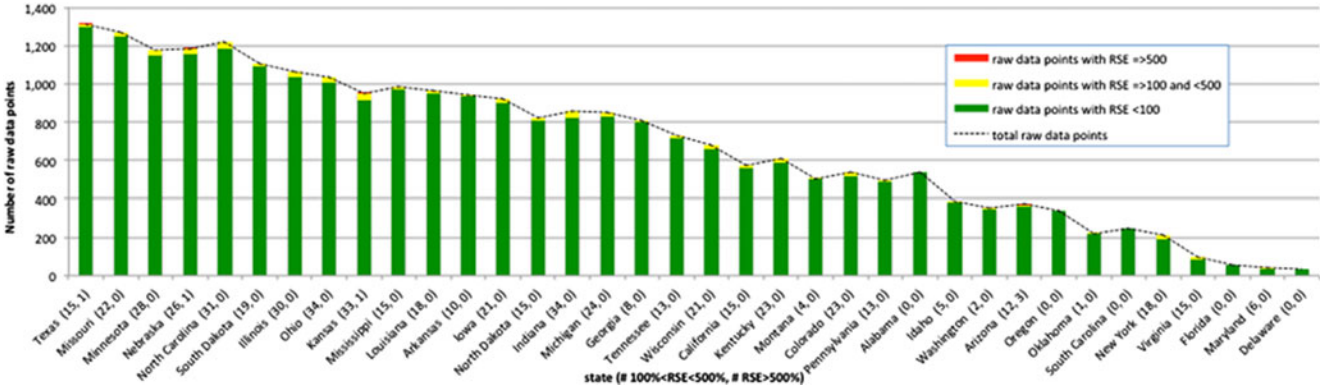

Fig. 2 ARMS raw data by state

RSEs for both fuels are 10 %, the means are statistically the same, and drawing the conclusion that the biofuel is superior is misleading. Thus, without knowledge of the error and sample sizes, the comparison of greenhouse gas emissions can be meaningless on the sole basis of the means, and as RSEs increase, it becomes less likely that conclusions regarding the difference in mean values for the sample statistic are statistically valid.

Moving from the hypothetical to actual LCA data, herein, we analyze the magnitude of variability (specifically, the sampling error) in unit process data representing field crop production. Field crop LCAs and related unit process data representing food, biofuel, and bioproducts are currently in high demand. In the USA, agricultural data relevant to LCA have been collected since 1810 (US Department of Agriculture 2011). Presently, the USDA's National Agricultural Statistics Service (NASS) conducts hundreds of surveys each year. Among the NASS surveys, a joint project with the USDA Economic Research Service (ERS), the Agricultural Resource Management Survey (ARMS[3]) provides field-level farm data that are particularly useful in the development of unit process data for LCA.

Specifically, ERS provides annual data summaries from the ARMS for field crops produced in 38 US states beginning in 1996 with only select crops surveyed each year: barley for malt and feed, corn, cotton, oats, peanuts, rice, sorghum, soybeans, durum wheat, other spring wheat, and winter wheat. Each ARMS crop–state–year combination (e.g., the production of soybeans in Iowa in 2006) covers seed use, irrigation technology and water use, tillage systems, nutrient and organic fertilizer (manure) use and management, crop residue management, and previous crop and pesticide use as defined by the ARMS variables.[4]

When the ARMS data are combined with NASS Quick Stats[5] data representing field crop production for each ARMS crop–state–year combination, the basis for an LCA unit process data flow is created. For example, the data for soybean production in Iowa in 2006 use the ARMS variables "Average seeding rate" (in pounds per acre) and "Planted acres" and are combined with NASS data representing the soybean production in Iowa in 2006 (in pounds) to estimate the seed use ultimately as kilograms of seeds per kilograms of soybeans produced in Iowa in 2006. To complete a field crop production unit process data set, additional information sources (e.g., data and documents from NASS, the Intergovernmental Panel on Climate Change, and more) are used to estimate a wide variety of activities and flows from and to nature.

Sommer et al. (1998) describe ARMS as a probability-based survey where each respondent represents a number of farms of similar size and type and the sample data are expanded using appropriate weights to represent operations at the state level. According to Kim et al. (2004), a delete-a-group jackknife variance estimator is used to describe how well a given estimate represents the population mean. "Jackknifing" is a resampling technique used to quantify bias and RSE by successively computing the mean, each time leaving out one or more groups of observations from the sample set. The RSE determined by a jackknife is a representation of the sensitivity on the estimate of the groups of samples used to produce that estimate and can be represented by an unbiased probability distribution such as a Student's *t*.

With the ARMS data, replicate weights are used to form a sample size of 15 or 30 replicate groups that are used for the jackknife estimation (15 prior to 2009 and 30 in 2009). Differences between the estimate and population mean result from nonsampling errors (e.g., related to questionnaire design or data processing) and sampling errors (e.g., related

[3] Data are available at http://www.ers.usda.gov/Data/ARMS/.

[4] See http://www.ers.usda.gov/Data/ARMS/Variables.htm for a list of ARMS variables.

[5] See http://quickstats.nass.usda.gov/.

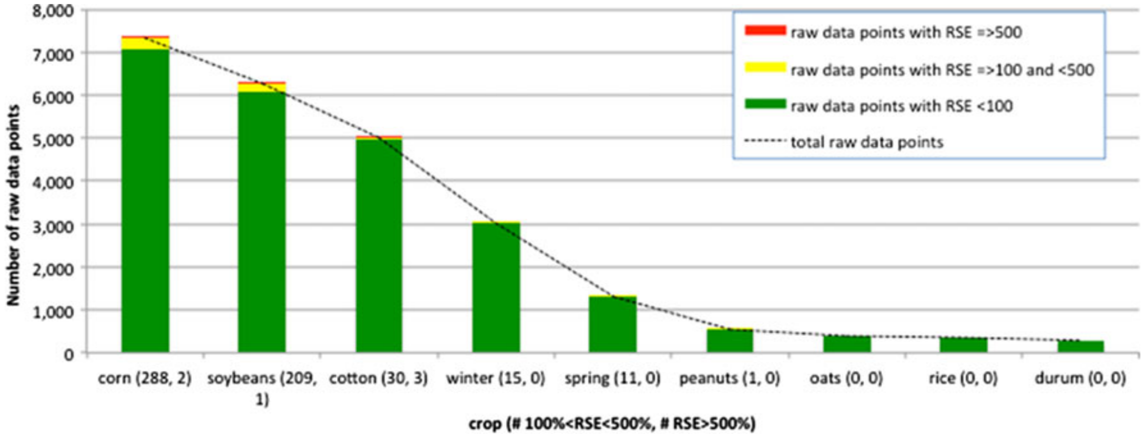

Fig. 3 ARMS raw data by crop

to sample selection, estimation, or nonresponse adjustments). Whereas nonsampling errors cannot be measured directly, sampling error is represented in ARMS as the jackknife RSE of the expected population mean. According to Dubman (2000) and Kim et al. (2004), RSE was chosen for the ARMS data as a measure of statistical reliability for two explicitly defined reasons: it is roughly equal to the expected value of the RSE of the population, and its measure of reliability is dependent on both the sample deviation and sample size. When calculations combine ARMS estimated means to estimate LCA exchange data, the ARMS jackknife RSEs are propagated based on the type of mathematical operation performed as described by Dieck (2007).

Given this, of interest here is to understand how the magnitude of the sampling error in the raw ARMS field crop data is propagated to sampling error in example unit process data. The overall intent is to begin a dialog, within the LCA practitioner community and among those using LCA results, concerning conclusive interpretations of comparative bioproduct LCA results.

2 Methods

ARMS data were used to prepare unit process data using parameterization (i.e., the presentation of data as formulas and the variables used) as they would be formatted for the European Reference Life Cycle Data System[6] according to the International Reference Life Cycle Data System (ILCD) data format and will be supported in the ecoinvent database[7] according to the EcoSpold v2 format. Because of the relatively small sample sizes of 15 or 30 used in the jackknife estimate of the ARMS means, a Student's *t* distribution is the appropriate representation of the probability density function (see Kim et al. 2004; Spiegel et al. 2009) and is thus used here. The RSE is used to construct a 95 % confidence interval for the estimated mean, assuming a *t* value of 2.145 for the 15 sample jackknives (at 14 degrees of freedom) and 2.045 for the 30 sample jackknives (at 29 degrees of freedom).

Given this, ARMS data were analyzed for 466 crop–state–year combinations (see Table S1 in the Electronic supplementary material) using 146 ARMS variables (see Table S2) in six categories: previous crop; till and seed; irrigation; synthetic fertilizer, lime, and nitrogen inhibitor; organic fertilizer; and pesticides. Of the possible 68,036 data points, values for 24,512 data points were available in ARMS with the remaining omitted as noncompliant with the NASS and ERS disclosure limitation practices, not available, or not applicable. The four units of measure for the variables were area (e.g., the planted or irrigated area or the area to which pesticide is applied), percent (e.g., the percent of the planted area treated with synthetic nitrogen fertilizer), depth (for the depth of irrigation water applied), and mass/area (e.g., mass of synthetic nitrogen fertilizer applied per treated area). All raw data (i.e., the farm data aggregated to the state level by ERS) can be downloaded directly from the ARMS website and note that the Supplemental electronic information has been intentionally left in the English units of measure of the raw ARMS data for the purpose of transparency.

Using the raw ARMS data with crop production data from NASS Quick Stats for each crop–state–year combination, 105 *unit process exchanges* and *interim calculations* were calculated. Unit process exchanges are flows that would appear in a unit process data set as calculated here, and interim calculations are data that require information beyond the ARMS and NASS data considered here to represent exchanges (e.g., the percent of nitrogen fertilizer that is ammonia, ammonia nitrate, urea, etc.). Noting that only a

[6] Available at http://lca.jrc.ec.europa.eu/lcainfohub/datasetArea.vm
[7] Available at http://www.ecoinvent.ch/

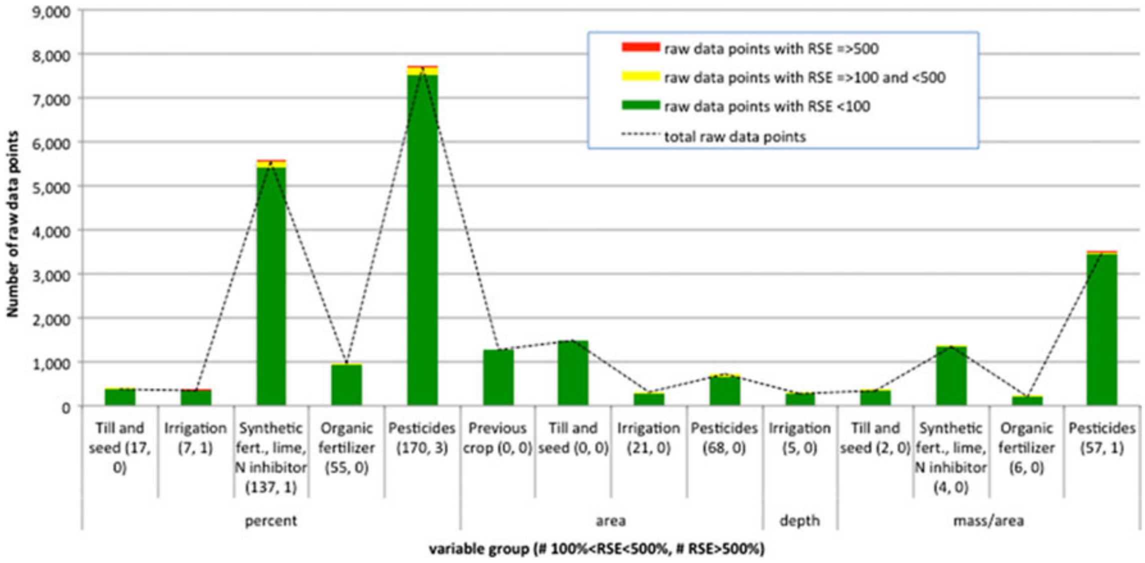

Fig. 4 ARMS raw data by variable group and units of measure

subset of the exchanges for the crop production unit process data area considered here (in fact representing only select technosphere flows), the parameterization of the exchanges and interim calculations represent three units of measure: area (e.g., on which organic fertilizer is injected/knifed in), mass (e.g., that applied as the active ingredient aryl triazolinone), and volume (e.g., of irrigation groundwater applied using pressure irrigation systems) (see Table S3 in the Electronic supplemental information). Only the parameters for the estimation of the exchanges or interim calculations are included here, with the parameterization of the RSE data described elsewhere (Cooper et al. 2011).

3 Results

3.1 Evaluation of the raw ARMS data

The RSE values of the ARMS variables investigated range from zero to over 1,600 %.[8] All were divided into three groups (Figs. 1, 2, 3, and 4): those with a RSE <100 %, those with a RSE between 100 and 500 %, and those with a RSE >500 % by year, crop, state, and ARMS variable group. Noting that the vast majority of the RSE values are <100 %, in particular the results should be viewed noting that the vast majority of the RSE values that are >100 % represent the synthetic nutrient and pesticide applications for corn and soybean production for which data were only collected in 2001 and 2002 (i.e., placing Figs. 1, 2, 3, and 4 within the context of Table S1 in the Electronic supplementary material). Also, there are only six data points with a RSE >500 % (three representing the production of cotton in Arizona in 1996, one representing the production of corn in Kansas in 2001, one representing the production of corn in Texas in 1999, and one representing the production of soybeans in Nebraska in 2002) covering nitrogen fertilizer application, pesticide application, and irrigation.

All data related to the production of durum wheat, rice, and oats have ARMS raw RSE values <100 %, and only one peanut-related data point had a RSE exceeding 100 %. Also, all data measured in area for previous crops and till and seed technology had RSE values <100 %. Finally, data collected outside of 2001–2002 are represented by data with RSE values <100 % for between 99 and 100 % of the data points.

Using Student's *t* distribution to represent the distribution of the raw ARMS data, it was found that many of the ARMS variables have 95 % confidence bounds that either fall below zero and/or, in the case of variables, measured as a percentage above 100 %. In fact, data with a 95 % confidence interval below zero represented 12 % of all raw data points, and percentage data with a 95 % confidence interval exceeding 100 % represented 7.4 % of all the raw data points. These phenomena dictate a need to be mindful of how the raw data are used to develop unit process data and ultimately how such data are combined into an inventory.

[8] The RSE value of 1,636 % for the crop–state–year combination cotton–Arizona–1996 representing the percent of nitrogen fertilizer broadcast with incorporation can be viewed at http://www.ers.usda.gov/Data/ARMS/app/default.aspx by selecting the survey "Crop production practices," the subject "Cotton," the filter by US/State "Arizona," from year "1996," and the report "Nutrient use by application method." The next two largest RSE values also represent cotton–Arizona–1996 followed by a RSE of 594 % for corn–Kansas–2001 representing the percent of insecticide acre treatments that were broadcast with incorporation.

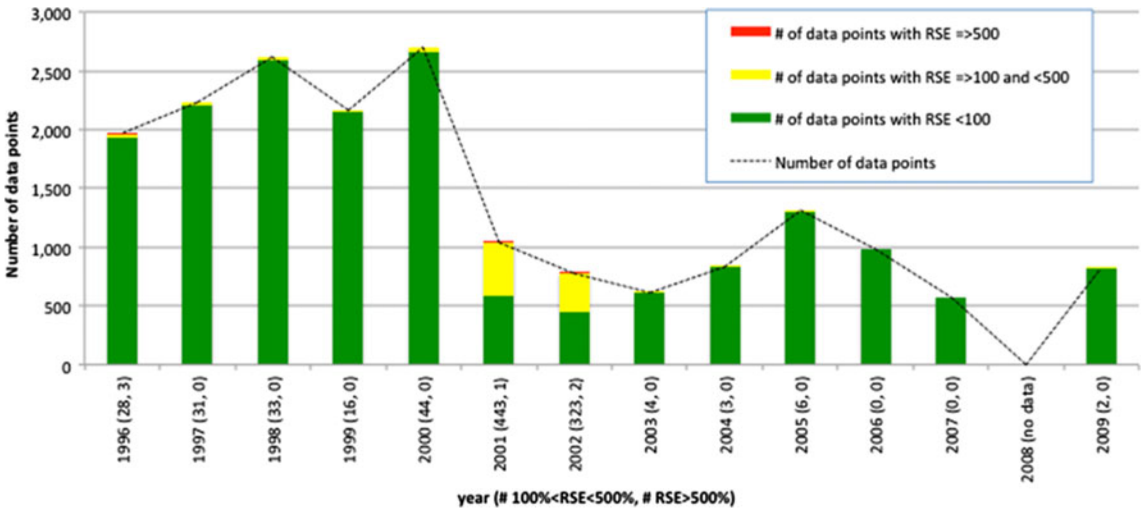

Fig. 5 Exchange and interim calculation data by year

3.2 Evaluation of the unit process data

Overall, 18,673 exchange and interim calculation data points were calculated, each with its respective RSE propagated from the raw data. Again, the vast majority of the RSE values are <100 % (Figs. 5, 6, 7, and 8) and range from 0 to over 1,600 % (see Tables S4–S6, Electronic supplementary material) with a greater portion of the data >100 % as single larger raw data RSE values used in multiple calculations. Again, the exchange and interim calculation data show a greater portion of the RSE >100 % for (a) data collected from 2001 to 2002, (b) data representing the production of corn and soybeans, and (c) data representing pesticide and synthetic applications; however, notably, the frequency of irrigation data with RSE >100 % is the largest among the exchange and interim calculation groups.

When the 95 % confidence bounds of the raw data fall below 0 and/or above 100 %, the characteristic is propagated to the un-normalized[9] and ultimately the normalized exchange and interim calculation data. For example, for the crop–state–year combination winter wheat–Texas–2009, the exchange data representing the area to which potassium fertilizer is applied are estimated to be 696,481 acres with a 95 % confidence interval from 541,298 to 851,674 acres. Of that area, 421,330 acres is estimated to broadcast potassium fertilizer with incorporation and 153,832 acres without incorporation (with the balance using an unspecified application method). However, the 95 % confidence intervals of the application methods are −21,559 to 864,218 and −78,662 to 386,325 acres for applications with and without incorporation, respectively. Thus, not only are the data wrongly inferring that the lower bounds are below 0 acres but also the upper bound of the area broadcast with incorporation exceeds the upper bound of the application area even before the area without incorporation is added to it. Thus, it is found that the probability density function for these data falls outside the actual limits for both the lower and upper tails.

Although the 95 % confidence interval does not include the full probability distribution function (which technically goes to ±infinity), here, the interval is used as an indication of how much of the exchange and interim calculation data fall outside actual limits. The result was that 20.3 % of the data points have a 95 % confidence interval lower bound less than 0 and 20.1 % are found to exceed the upper limit of the 95 % confidence interval of the interim calculation for which they are based.

4 Discussion

The research presented here was motivated by an interest in understanding the magnitude of sampling error in crop production unit process data for LCA within the context of conclusive interpretations of comparative bioproduct LCA results. Towards this, select exchanges from the technosphere and related interim calculations were developed from the ARMS data. With RSE values ranging from 0 % to greater than 1,600 %, the least precision was found in data collected between 2001 and 2002, in the production of corn and soybeans, and in synthetic and pesticide applications and irrigation data. The highest precision was seen in data representing the production of durum wheat, rice, oats, and peanuts and in data representing previous crops and till and seed technology use.

High RSE values arise from the RSE representing a biased distribution, a jackknife estimate being nearly zero,

[9] As in not divided by production (named PROD in Table S3 of the Supplemental electronic information)

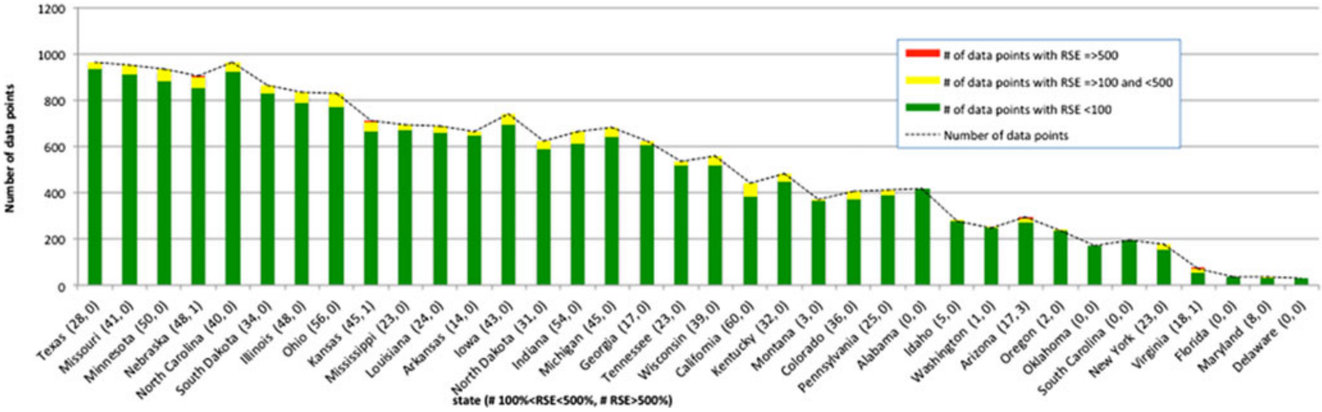

Fig. 6 Exchange and interim calculation data by state

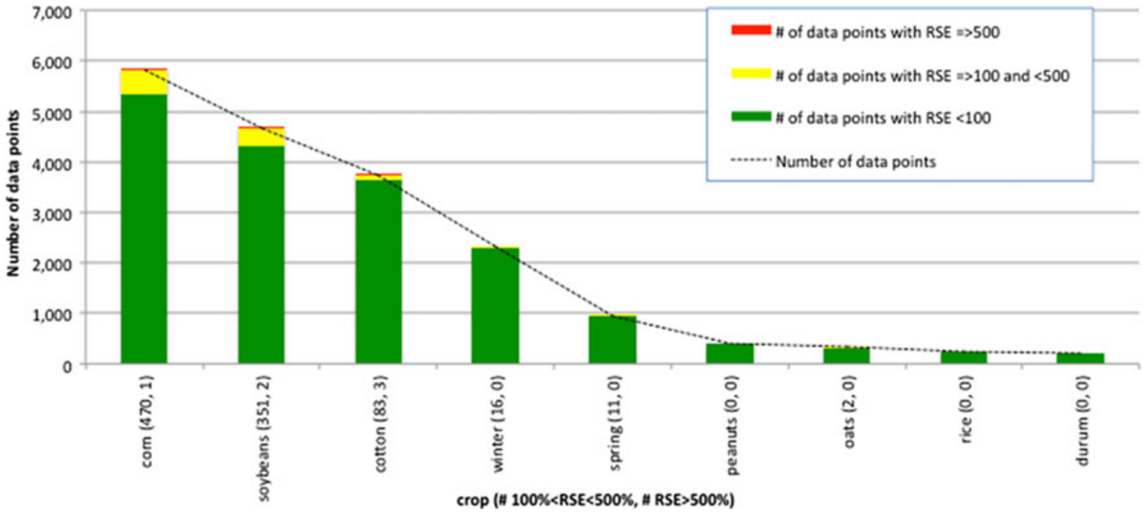

Fig. 7 Exchange and interim calculation data by crop

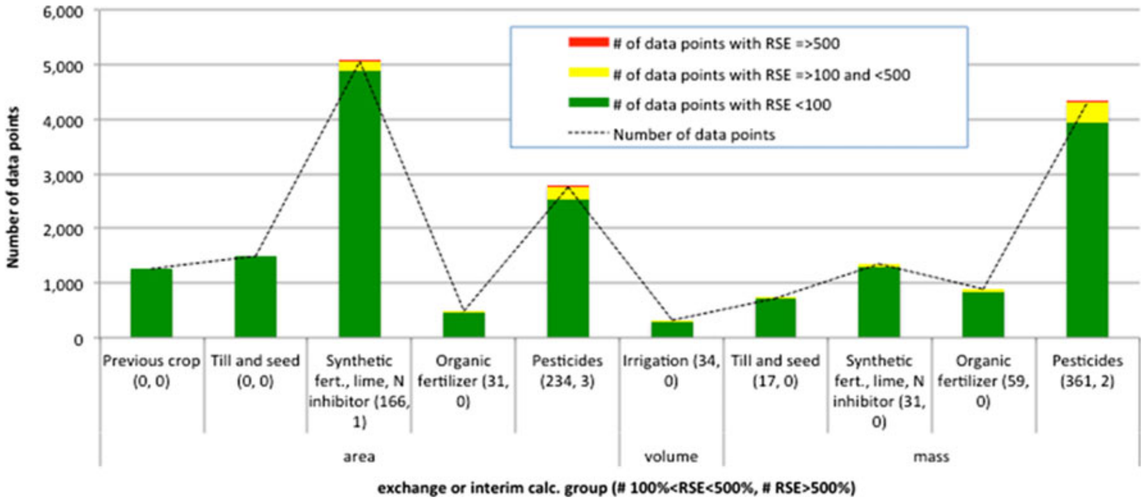

Fig. 8 Exchange and interim calculation data by variable group and units of measure

or error propagation using low-precision data. Sommer et al. (1998) note that the higher the ARMS RSE, the less well the estimate represents individual items in the delete-a-group jackknife. They also note that the ARMS data are also influenced by nonsampling errors and that efforts are taken to minimize them. Given this, Kim et al. (2004) note that the magnitude of the ARMS data bias is unknown and that the reliability of an ARMS estimate cannot be tested when there is no knowledge of the distribution because the population variance is unknown—i.e., the reliability test for the sample mean can be made only under the normality assumption and leading to the use of Student's *t* distribution due to the low number of jackknife samples. Also, many of the ARMS variables describe positive definite parameters, depth of irrigation water or acres of herbicide applied, as examples. Unless negative weights are applied to groups during the jackknife, an estimate mean with a value nearly zero should not be sufficient to produce an RSE greater than 100 % for a positive definite or semidefinite, unbiased parameter, noting that none of the jackknife samples should be negative for positive semidefinite parameters. Within this context, guidance can be taken from ARMS in which data with a RSE >25 % are deemed statistically unreliable, for example due to low sample size and/or a high sampling error. The unit process data prepared from this work will also mark such data in a comment data field.

Further, here it is found that a portion of the data is represented by a 95 % confidence interval that falls outside actual limits. Confidence intervals beyond physical bounds are entirely possible due to the high standard errors that are a consequence of using a jackknife on subsets of data for which the weights are not calibrated explicitly and a low presence of certain practices. Such data essentially represent a truncated Student's *t* distribution, which when interpreting an inventory, e.g., using Monte Carlo simulation, should only be sampled within the appropriate bounds. With the advent of parameterization in LCA data formats, which provides the opportunity to include raw data and the formulas that use them within a unit process data set, the raw percentage data can be kept within appropriate bounds while still maintaining the distribution of interest, as described by Cooper et al. (2011).

At high levels of sampling error such as those described here, comparisons of LCA bioproduct results must be made with caution and must be tested to ensure mean values are different to a desired level of significance. As the use of LCA is growing in decisions being made pursuant to public policy, law, and product comparisons, the need for uncertainty data grows as well. Emerging data formats such as ILCD and EcoSpold v2 that allow parameterization in a way that uncertainty can be propagated from raw data to exchange provides another important component of a move towards improved LCA data and improved LCAs.

All data are expected to be available through the USDA LCA Digital Commons (at http://www.openlca.org/index.html) early in 2012.

Acknowledgments The US Department of Agriculture (USDA) National Agricultural Library (agreement number 58-8201-0-149) funded this research. The views expressed are those of the authors and should not be attributed to the Economic Research Service or the US Department of Agriculture. The views expressed in this article are those of the authors and do not necessarily reflect those of the U.S. Department of Agriculture or Economic Research Service.

References

Cooper JS, Noon M, Kahn E (2011) Parameterization in life cycle assessment inventory data: review of current use and the representation of uncertainty. Int J Life Cycle Assess. doi:10.1007/s11367-012-0411-1

Dieck RH (2007) Measurement uncertainty—methods and applications, 4th edn. International Society of Automation, Research Triangle Park

Dubman RW (2000) Variance estimation with USDA's farm costs and returns surveys and agricultural resource management study surveys. US Department of Agriculture, Economic Research Service Resource Economics Division

Kim CS, Hallahan C, Lindamood W, Schaible G, Payne J (2004) A note on the reliability tests of estimates from ARMS data. Agr Resource Econ Rev 33(2):293–297

Lloyd SM, Ries R (2007) Characterizing, propagating, and analyzing uncertainty in life-cycle assessment. A survey of quantitative approaches. J Ind Ecol 11(1):161–179

Sommer JE, Hoppe RA, Green RC, Korb PJ (1998) Structural and financial characteristics of US farms, 1995: 20th Annual Family Farm Report to Congress. Retrieved from http://www.ers.usda.gov/publications/aib746/. Accessed 15 Feb 2012

Spiegel MR, Schiller JJ, Srinivasan RA, Alu R (2009) Schaum's outlines—probability and statistics, 3rd edn. McGraw-Hill, New York

US Department of Agriculture (2011) 2007 Census of agriculture: history volume 2 subject series Part 7. National Agricultural Statistics Service. Retrieved from: http://www.agcensus.usda.gov/Publications/2007/Full_Report/2007%20History%20of%20the%20Census4-7(f).pdf. Accessed 15 Feb 2012

Weidema BP, Wesnæs MS (1996) Data quality management for life cycle inventories—an example of using data quality indicators. J Cleaner Prod 4(3–4):167–174

Weidema BP, Bauer C, Hischier R, Mutel C, Nemecek T, Vadenbo CO, Wernet G (2011) Overview and methodology: data quality guideline for the ecoinvent database version 3 (final draft_revision 1) ecoinvent report no. 1(v3), http://www.ecoinvent.org/fileadmin/documents/en/ecoinvent_v3_elements/01_DataQualityGuideline_FinalDraft_rev1.pdf. Accessed 15 Feb 2012